



# **Vo Declaration Exhibit J**


# EXHIBIT H

Fair-Use Lib: 400B+ New Clean Tokens (+325B more to come)

 Nikolay Bashlykov and 4 others · Last edited July 28, 2023 · 3 minute read



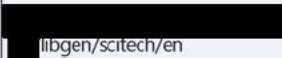
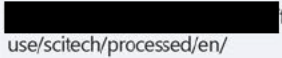
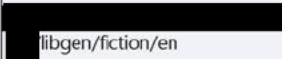
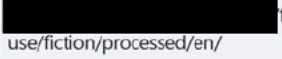




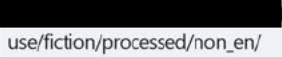
 Save ...

# Fair-Use Lib: 400B+ New Clean Tokens (+325B more to come)

**TL;DR;** We have collected a new **400B** dataset of high-quality tokens on almost every possible subject from STEM and fiction books to cooking, gardening and historic books. This dataset is **5x larger** than the full  dataset, while having the **comparable data quality**. See more details in the lab-notebook.

by: Lukas Blecher, Peter Albert, Angela Fan, Naman Goyal, Nikolay Bashlykov

## Location

Part			Tokens
EN Scientific Books	 libgen/scitech/en	 fair-use/scitech/processed/en/	220B
EN Fiction Books	 libgen/fiction/en	 fair-use/fiction/processed/en/	110B
EN Scientific Articles	TBD	TBD	TBD
non-EN Scientific Books	 ed/libgen/scitech/non_en	 fair-use/scitech/processed/non_en/	15B
non-EN Fiction Books	 fair_llm/data  libgen/fiction/non_en	 fair-use/fiction/processed/non_en/	55B

## Data curation steps

We have lost roughly 5-7% of corrupted files (broken PDFs/EPUBs) and additional 1.2% after post-processing filtering. So overall we have preserved ~90% of cleaned tokens.

1. **Data Sourcing:** downloaded **1.6M Scientific, 1.1M Fiction books and 80M scientific articles**. We pre-filtered PDF/EPUB using metadata provided by the library.

**2. Data Parsing:**

- a. PDF parsing is done using OCR Nougat
- b. EPUB parsing is done by a wrapper around html\_to\_latex.py used for Common Crawl html parsing

**3. Data Cleaning (filtering):**

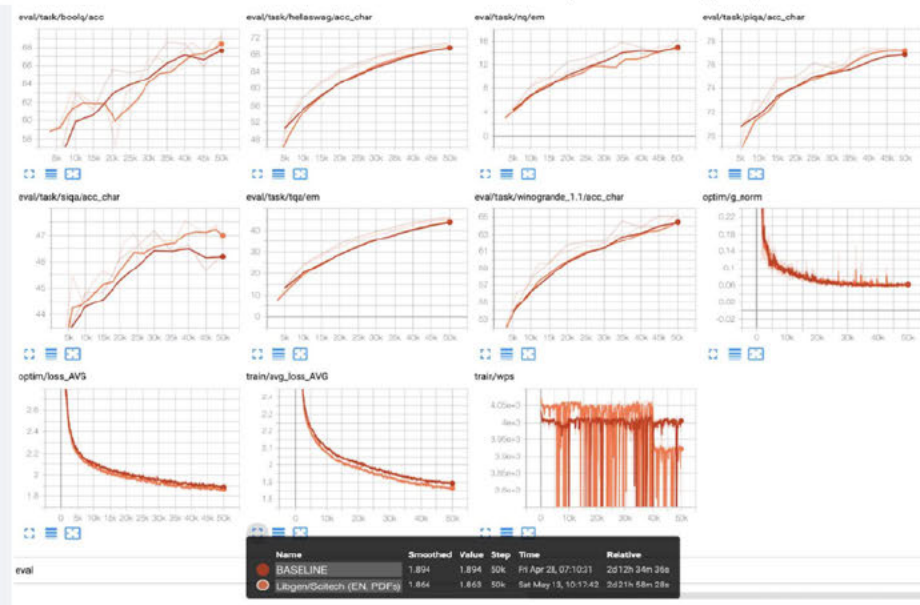
- a. Remove artifacts after OCR parsing (for PDFs only)
  - i. Clean duplicated sentences within the text
  - ii. Fix markdown for tables and lists
  - iii. Remove small artifacts: hallucinations (references, headers), title numbering, repetitions
- b. General post-processing steps/filtering
  - i. Remove references lists, copyright sections ("ISBN", "Copyright", "©", "All rights reserved"),
  - ii. Remove boilerplate ("ISBN", "Copyright", "©", "All rights reserved")
  - iii. Remove short books (line count < 50)
  - iv. Remove corrupted lines (Non-empty lines have less than 20 characters avg length, numeric fraction of characters > 10%, line longer than 50k characters)
  - v. Main language (top probability after fasttext language classifier) less than 0.5

**Ablation results**

No red flags observed. Some improvement on siqa and boolq.

9/4/24, 6:16 PM

Fair-Use Lib: 400B+ New Clean Tokens (+325B more to come) | Workplace



### Comparison

- Scientific Books (12x larger): 300k [REDACTED] VS 3.7M Fair-Use Lib
- Scientific Articles (20x larger): 4M [REDACTED] VS 82M Fair-Use Lib
- Fiction Books: 0 [REDACTED] VS 2.7M Fair-Use Lib

Random samples of parsed data:

#### Introduction

Few remodeling jobs have a greater impact on the livability and value of your home than replacing, adding or repairing windows. There are many paybacks: improved home value, better security, lower energy costs, increased light and ventilation, and added curb appeal. Best of all, if you have moderate tool skills, window projects are well within your reach, saving you considerable costs over having the work done by contractors.

Unlike the larger, more expensive volumes, this book doesn't include background information on cutting holes in walls or expanding living areas. What it does do is give you step-by-step information on how to change the look of your home by installing different types of windows.

Want to add some indoor greenery? Check out the Garden Window on page 22. Need to add some privacy to a bathroom while still allowing some natural light? Turn to page 26 to see how to install a glass block window. Looking for a way to add some light to a dark interior room? The skylight project on page 40 may be just what you're looking for.

And if you aren't ready to replace a window just yet, there are steps that can be taken to repair or improve operation of your current windows. There's even a section on inexpensive ways to reduce energy loss through your windows or patio doors.

Window replacement can be very expensive, but with *Here's How: Windows*, you may find that you're able to save thousands of dollars doing it yourself, and giving your home an upgrade.

#### image Window Styles

The following pages contain examples of some of the types of windows you may consider for your home.

Your imagination may lead you to other options and combinations of options.

Casement windows pivot on hinges mounted on the side. They have a contemporary look and offer good ventilation. Whether your window has exposed or concealed sash locks, casements have a reputation for weather-tight construction.

Double-hung windows slide up and down and have a traditional appearance. The newer-style, spring-mounted operating mechanism is virtually trouble-free. The dividers (muntins) may divide individual panes of glass or snap on for decoration.

#### The MAN WHO TASTED ASHES

**\*\*Algis Budrys\*\***

THE CAR HED stolen was a beautifully groomed thing: all polished lacquer and chrome, with almost brand-new dual tread whitewall tires on the nicked wire wheels. But the transmission was bad, the brake drums scraped, and there was a short circuit in the wiring somewhere, so that he had to keep over sixty miles per hour or the generator would not charge at all. He would have stolen another one if he could, but he had got onto the turnpike before he realized just how unreliable this one was. If he changed cars at a restaurant, it would be reported and the police would stop him when he tried to leave the turnpike.

No, he was trapped with what he had. Hunched over the wheel of his roaring cage, the yellowish headlights reflecting white from the lane markers, Redfern swept his eyes systematically over the instruments: ammeter, fuel gauge, oil pressure, water temperature, speedometer, odometer. He thought of himself as doing it systematically, every ten minutes, like a professionally trained driver. Actually, he was dividing his attention almost equally between the road and the odometer. A hundred and ten miles covered, seventy miles to go, ninety minutes before the ship was due to take off, with or without him, average speed required: 42.62, approx.; round off to allow for stopping the car at the exit toll booth, for covering two miles of back roads, for leaving the car and running an unknown distance across a weed-grown field to the ship's airlock—they would take off on schedule with him six inches from the slamming airlock door; they would not stay themselves a microsecond to accommodate him—say fifty miles per hour, average. Then allow for speedometer error. Say fifty-five miles per hour, indicated, average. Allow for odometer error. Say sixty miles per hour, indicated, average. Allow for unforeseen delays. Sixty-five miles per hour.

Redfern's foot trembled on the accelerator pedal. His thigh ached from hours of unremitting pressure. His car flashed by signboards, wove continually around immense trailer trucks in the slow lane. His mind raced to keep up with the changing figures on the odometer. He wished he weren't feeling a slight miss in the engine whenever he eased up on the accelerator. He cursed the car's owner for his false-front prodigality with wax and whitewalls.

He looked at his watch again. Four in the morning. He turned the radio on, ignoring his fear that something else might happen to the car's wiring.

## Next Steps

9/4/24, 6:16 PM

Fair-Use Lib: 400B+ New Clean Tokens (+325B more to come) | Workplace

- Finish downloading & Parse Scientific Articles part (80M articles) -> **additional ~300B tokens**
- Download & Parse other languages. We would need to update the OCR parsing library to work with non-EN languages (top-5 languages). -> **additional ~150B tokens**



5

2 comments



Like



Comment



Share

Most relevant ▾

**Dieuwke Hupkes**

Wow this is awesome! Somehow just seeing this now.

50w Like Reply

**Nikolay Bashlykov** **Dieuwke Hupkes** yes, here is the full report with 650B tokens:<https://docs.google.com/> 

Connect Google Drive account to see this preview

50w Like

Reply



Write a comment...

